Journal of Multivariate Analysis 101 (2010) 824–838

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Tests for multiple regression based on simplicial depth

Robin Wellmann *, Christine H. Müller

Department of Mathematics, University of Kassel, D-34109 Kassel, Germany

ARTICLE INFO

Article history:
Received 4 August 2008
Available online 6 January 2010

AMS subject classification:
primary 62G05
62G10
secondary 62J05
62J12
62G20

Keywords:
Degenerated U-statistic
Distribution free tests
Multiple regression
Outlier robustness
Regression depth
Simplicial depth
Spectral decomposition
Shape analysis

ABSTRACT

A general approach for developing distribution free tests for general linear models based on simplicial depth is applied to multiple regression. The tests are based on the asymptotic distribution of the simplicial regression depth, which depends only on the distribution law of the vector product of regressor variables. Based on this formula, the spectral decomposition and thus the asymptotic distribution is derived for multiple regression through the origin and multiple regression with Cauchy distributed explanatory variables. The errors may be heteroscedastic and the concrete form of the error distribution does not need to be known. Moreover, the asymptotic distribution for multiple regression with intercept does not depend on the location and scale of the explanatory variables. A simulation study suggests that the tests can be applied also to normal distributed explanatory variables. An application on multiple regression for shape analysis of fishes demonstrates the applicability of the new tests and in particular their outlier robustness.

1. Introduction

Liu [1,2] used the half space depth of Tukey [3] to define simplicial depth of a multivariate location parameter $\theta \in \Theta = \mathbb{R}^q$ in a sample $z_1, \ldots, z_N \in \mathbb{R}^q$ as

$$d_S(\theta, (z_1, \ldots, z_N)) = \binom{N}{q+1}^{-1} \sum_{1 \le n_1 < n_2 < \cdots < n_{q+1} \le N} \mathbb{I}\{d(\theta, (z_{n_1}, \ldots, z_{n_{q+1}})) > 0\}, \tag{1}$$

where d is the half space depth of Tukey and $\mathbb{I}$ denotes the indicator function. This depth counts the simplices spanned by $q+1$ data points which are containing the parameter θ. Since Tukey [3], several other depth notions were introduced. Each of them can be used as depth d in (1) leading to several different simplicial depth notions. Several depth notions can be obtained from the book of Mosler [4] and the references therein. If d is the regression depth of Rousseeuw and Hubert [5], then d_S is called simplicial regression depth. General concepts of depth were introduced and discussed by Zuo and Serfling [6,7] and Mizera [8]. Mizera [8], in particular generalized the regression depth of [5] by based on the quality functions instead of squared residuals. This approach makes it possible to define the depth of a parameter value with respect to given observations in various statistical models via general quality functions. Appropriate quality functions are in particular likelihood functions as studied by Mizera and Müller [9] for the location − scale model and by Müller [10] for generalized linear models.

* Corresponding author.
E-mail address: r.wellmann@uni-hohenheim.de (R. Wellmann).

0047-259X/$ – see front matter
doi:10.1016/j.jmva.2009.12.008

Any concept of data depth can be used to generalize the notion of ranks and to derive distribution free tests by generalizing Wilcoxon's rank sum test. Nevertheless only few papers deal with tests based on data depth. Liu [11] and Liu and Singh [12] proposed distribution free multivariate rank tests based on depth notions. While they proved its asymptotic normality for special depth notions and special dimensions, Zuo and He [13] were able to extend these results to general situations. Other distribution free tests are provided by the concept of ranks and signs based on the multivariate Oja median (see [14]). For an overview of these methods, see [15]. However all these approaches provide only tests for multivariate data and do not concern regression models. Bai and He [16] derived the asymptotic distribution of the maximum regression depth estimator. However, this asymptotic distribution is given implicitly so that it is not convenient for testing. Tests for regression based on depth notions were only derived by Van Aelst et al. [27], Müller [10] and Wellmann et al. [17]. Van Aelst et al. [27] even derived an exact test based on the regression depth of Rousseeuw and Hubert [5] but did it only for linear regression. For multiple regression his approach can be used based on simulated quantiles.

Müller [10] and Wellmann et al. [17] used the fact that any simplicial depth is a U-statistic with kernel function

$$\psi_\theta(z_{n_1}, \ldots, z_{n_{q+1}}) = \mathbb{I}\{d(\theta, (z_{n_1}, \ldots, z_{n_{q+1}})) > 0\}.$$

For U-statistics the asymptotic distribution is known (see e.g. [18,28]). However, the U-statistic is degenerated for most simplicial depth notions so that the spectral decomposition of the conditional expectation

$$\psi_\theta^2(z_1, z_2) := E_\theta(\psi_\theta(Z_1, \ldots, Z_{q+1})|Z_1 = z_1, Z_2 = z_2) - E_\theta(\psi_\theta(Z_1, \ldots, Z_{q+1})) \tag{2}$$

is needed to derive the asymptotic distribution. But as soon as the spectral decomposition of (2) is known, asymptotic tests can be derived for any hypothesis of the form $H_0 : \theta \in \Theta_0$ where Θ_0 is an arbitrary subset of the parameter space Θ, provided that the asymptotic distribution does not depend on the unknown parameter. These tests are based on the test statistic $T(z_1, \ldots, z_N) := \sup_{\theta\in\Theta_0} T_\theta(z_1, \ldots, z_N)$, where $T_\theta(z_1, \ldots, z_N)$ is defined as

$$T_\theta(z_1, \ldots, z_N) := N(d_S(\theta, (z_1, \ldots, z_N)) - \mu_\theta) \tag{3}$$

with $\mu_\theta = E_\theta(\psi_\theta(Z_1, \ldots, Z_{q+1}))$. The hypothesis H_0 is rejected if $T(z_1, \ldots, z_N)$ is smaller than the α-quantile of the asymptotic distribution of $T_\theta(Z_1, \ldots, Z_N)$, see [10,17].

The spectral decomposition of (2) was derived by Müller [10] for linear and quadratic regression by solving differential equations. Wellmann et al. [17] extended this result to polynomial regression with polynomials of arbitrary degree by proving a general formula of (2) and then specifying the general formula for polynomial regression so that the spectral decomposition can be found by Fourier series representation.

The general formula can be specified also for multiple regression so that a spectral decomposition of (2) can be derived for this case as well. This is shown in this paper.

In Section 2, general assumptions and definitions and in particular the general formula are given. In Section 3 the general formula is specified for multiple regression through the origin. Based on the specified formula the spectral decomposition is derived, which is given by spherical functions and eigenvalues depending on Gegenbauer functions.

The asymptotic distribution for multiple regression with intercept, where the regressors have Cauchy distribution, is given in Section 4. This model is traced back to multiple regression through the origin by multiplying the regressors and the dependent variables with additional random variables S_n. A simulation study, which is presented in Section 5, suggests that the test for multiple regression with intercept can be applied also to normal distributed explanatory variables. Another simulation study in Section 5 provides a power comparison with the classical F-test and the regression depth test of [27].

Section 6 presents some applications on tests in multiple regression through the origin with two explanatory variables in the shape analysis of fishes. These examples in particular show that the new tests possess high outlier robustness. All proofs are given in Section 7.

2. The general case

We assume a statistical model for i.i.d. random variables $Z_1, \ldots, Z_N$ with values in $\mathcal{Z} \subset \mathbb{R}^p$, $p \geq 1$ and parameter space $\Theta = \mathbb{R}^q$. We choose functions $h : \mathcal{Z} \to \mathbb{R}$ and $v : \mathcal{Z} \to \mathbb{R}^q$ and call

$Y_n := h(Z_n)$ the dependent variable,
$X_n := v(Z_n)$ the regressor, and
$S_n(\theta) := \operatorname{sign}(Y_n - X_n^T\theta), \quad \theta \in \mathbb{R}^q,$ the sign of the residual.

We assume that for all $\theta \in \Theta$:

$$\begin{aligned} &\bullet\ P_\theta(S_1(\theta) = 1|X_1) \equiv \frac{1}{2} \quad \text{a.s.,} \\ &\bullet\ P_\theta(S_1(\theta) = 0|X_1) \equiv 0 \quad \text{a.s., and} \\ &\bullet\ P_\theta(X_1, \ldots, X_q \text{ are linearly dependent}) = 0. \end{aligned} \tag{4}$$

The last two conditions of (4) are easily satisfied for example by continuous distributions. Depending on the distribution of Z_n, the first condition can be satisfied by appropriate transformations v and h. The first condition in particular implies that the true regression function is in the center of the data, which means that the median of the residuals is zero.

We denote random variables by capital letters and realizations by small letters. The depth of $\theta \in \Theta$ for observations $z = (z_1, \ldots, z_N)$ is given by

$$d_T(\theta, z) = \min_{u \neq 0} \#\{n : s_n(\theta) u^T v(z_n) \geq 0\}.$$

This depth coincides with the regression depth of [5] and with Definition 2 from [17], if the quality functions $g_{z_n}(\theta) = -(h(z_n) - v(z_n)^T\theta)^2$ are used. It is a tangent depth in the sense of Mizera [8]. As in [17], we work with a harmonized depth to improve the power of tests, i.e. we use

$$\psi_\theta(z_1, \ldots, z_{q+1}) = \begin{cases} d_T(\theta, (z_1, \ldots, z_{q+1})), & \text{if } s_n(\theta) \neq 0 \text{ for } n = 1, \ldots, q+1 \\ 0, & \text{otherwise,} \end{cases}$$

and define the simplicial depth as

$$d_S(\theta, z) = \binom{N}{q+1}^{-1} \sum_{1 \leq n_1 < n_2 < \cdots < n_{q+1} \leq N} \psi_\theta(z_{n_1}, \ldots, z_{n_{q+1}}).$$

Under the assumptions (4) we have

$$\mu_\theta = E_\theta(\psi_\theta(Z_1, \ldots, Z_{q+1}) | Z_1 = z_1) = \frac{1}{2^q}$$

(see [17]), so that $d_S(\theta, z)$ is a degenerated U-statistic. Hence the spectral decomposition of (2) is needed. This can be derived by the general formula of (2) given in [17]. It has the form

$$\psi_\theta^2(z_1, z_2) = \frac{s_1(\theta) s_2(\theta)}{2^{q-1}} \mathcal{K}(x_1, x_2),$$

with

$$\mathcal{K}(x_1, x_2) := P_\theta(x_1^T W x_2^T W < 0) - \frac{1}{2}, \quad \text{for } x_1, x_2 \in \mathbb{R}^q, \tag{5}$$

where $W := X_3 \times \cdots \times X_{q+1}$ is the vector product of $X_3, \ldots, X_{q+1}$. Hence only the spectral decomposition of the kernel $\mathcal{K}$ is needed. As soon as $\mathcal{K}$ does not depend on θ, which is the case for usual regression problems, the asymptotic distribution is independent of θ and the test given by (3) with $\mu_\theta = \frac{1}{2^q}$ can be used.

3. Multiple regression through the origin

Assuming a model for multiple regression through the origin,

$$Y_n = \theta_1 X_{n,1} + \cdots + \theta_q X_{n,q} + E_n = X_n^T \theta + E_n$$

we suppose that (4) holds and that there is an invertible matrix $A \in \mathbb{R}^{q \times q}$, such that $\frac{1}{\|AX_n\|} AX_n$ is uniformly distributed on the unit sphere. This is in particular the case, if X_n has an elliptical distribution such as the multivariate normal distribution with mean zero. In order to derive the asymptotic distribution of the simplicial depth for this regression model, we have to simplify the kernel function $\mathcal{K}$ given by Eq. (5). By using that with $\frac{1}{\|AX_3\|} AX_3, \ldots, \frac{1}{\|AX_{q+1}\|} AX_{q+1}$ also the normalized vector product is uniformly distributed on the unit sphere, we obtain the following proposition.

Proposition 1. *For all $x_1, x_2 \in \mathbb{R}^q \setminus \{0\}$ we have*

$$\mathcal{K}(x_1, x_2) = \frac{1}{\pi} \arccos\left(\left\langle \frac{Ax_1}{\|Ax_1\|}, \frac{Ax_2}{\|Ax_2\|} \right\rangle\right) - \frac{1}{2}.$$

The value $\mathcal{K}(x_1, x_2)$ depends only on the angle between Ax_1 and Ax_2. Thus, the required eigenvalues of the integral operator

$$T_\mathcal{K} : \mathbb{L}_2(P^{X_1}) \to \mathbb{L}_2(P^{X_1}) \quad \text{with } T_\mathcal{K} f(s) = \int \mathcal{K}(s, t) f(t) \mathrm{d}P^{X_1}(t)$$

depend on Gegenbauer functions (see [19]). The general formulas for the eigenvalues are obtained from the next proposition.

Proposition 2. *Let $S \subset \mathbb{R}^q$ be the unit sphere, where $q \geq 2$. Let $K : S \times S \to \mathbb{R}$, $K(s,t) := \frac{1}{\pi}\arccos(\langle s,t\rangle) - \frac{1}{2}$. The values*

$$\lambda_0 := 0$$
$$\lambda_p := -\frac{1}{2}\tau_q \left(\frac{\Gamma\left(\frac{q}{2}\right)\Gamma\left(\frac{p}{2}\right)}{\Gamma\left(\frac{q}{2}+\frac{p}{2}\right)} \frac{\sin\left(\frac{p}{2}\pi\right)}{\pi} \right)^2 \quad \text{for } p \in \mathbb{N}$$

are the eigenvalues of the integral operator T_K, where $\tau_q = 2\frac{\pi^{\frac{q}{2}}}{\Gamma\left(\frac{q}{2}\right)}$ is the $(q-1)$-dimensional volume of the sphere. For $p \in \mathbb{N}$, the corresponding eigenfunctions with respect to the uniform measure υ on S with $\upsilon(S) = \tau_q$ are the orthogonalized and normalized spherical functions $S^{(n)}_{(p,1)}, \ldots, S^{(n)}_{(p,u_p)}$ of degree p, where $n := q-2$. By Fenyö and Stolle [19] we have $u_p = \frac{(p+n-1)!}{p!n!}(2p+n)$.

Let $\left(S^{(q-2)}_{(p,k)}\right)_{(p,k)\in I}$ be the family of orthogonalized and normalized spherical functions from Proposition 2 with $I := \{(p,k) \in \mathbb{N}^2 : k \leq u_p\}$ and for $j \in I$ let $\varphi_j(x) := \sqrt{\tau_q} S^{(q-2)}_j\left(\frac{1}{\|Ax\|}Ax\right)$.

Because of $\frac{1}{\|AX_1\|}AX_1 \sim \frac{1}{\tau_q}\upsilon$, we obtain for all $i,j \in I$:

$$\begin{aligned}
\int \varphi_i\varphi_j \mathrm{d}P^{X_1} &= \int \sqrt{\tau_q} S^{(q-2)}_i\left(\frac{1}{\|Ax\|}Ax\right)\sqrt{\tau_q} S^{(q-2)}_j\left(\frac{1}{\|Ax\|}Ax\right) P^{X_1}(\mathrm{d}x) \\
&= \tau_q \int S^{(q-2)}_i\left(\frac{1}{\|AX_1\|}AX_1\right) S^{(q-2)}_j\left(\frac{1}{\|AX_1\|}AX_1\right)\mathrm{d}P \\
&= \tau_q \int S^{(q-2)}_i S^{(q-2)}_j \mathrm{d}P^{\frac{1}{\|AX_1\|}AX_1} \\
&= \frac{\tau_q}{\tau_q}\int S^{(q-2)}_i S^{(q-2)}_j \mathrm{d}\upsilon.
\end{aligned}$$

Hence, $(\varphi_j)_{j\in I}$ is an ONS in $\mathbb{L}_2\left(P^{X_1}\right)$. From the previous propositions we conclude, that in $\mathbb{L}_2\left(P^{X_1}\otimes P^{X_1}\right)$ we have:

$$\begin{aligned}
\mathcal{K}(x_1,x_2) &= \frac{1}{\pi}\arccos\left(\left\langle \frac{Ax_1}{\|Ax_1\|}, \frac{Ax_2}{\|Ax_2\|}\right\rangle\right) - \frac{1}{2} \\
&= \sum_{(p,k)\in I} \lambda_p S^{(q-2)}_{(p,k)}\left(\frac{1}{\|Ax_1\|}Ax_1\right) S^{(q-2)}_{(p,k)}\left(\frac{1}{\|Ax_2\|}Ax_2\right) \\
&= \sum_{(p,k)\in I} \frac{\lambda_p}{\tau_q}\sqrt{\tau_q} S^{(q-2)}_{(p,k)}\left(\frac{1}{\|Ax_1\|}Ax_1\right)\sqrt{\tau_q} S^{(q-2)}_{(p,k)}\left(\frac{1}{\|Ax_2\|}Ax_2\right) \\
&= \sum_{(p,k)\in I} \frac{\lambda_p}{\tau_q}\varphi_{(p,k)}(x_1)\varphi_{(p,k)}(x_2).
\end{aligned}$$

Hence with the Hoeffding decomposition of U-statistics (see e.g. [18], p. 79, 80, 90, [28], p. 650), we immediately get the next theorem:

Theorem 1. *Suppose, that there is an invertible matrix $A \in \mathbb{R}^{q\times q}$ with $q \geq 2$, such that $\frac{1}{\|AX_n\|}AX_n$ is uniformly distributed on the unit sphere and suppose that assumption (4) holds. Let $\lambda_1, \lambda_2, \ldots$ and $u_1, u_2, \ldots$ be as in the previous proposition.*

Then there are i.i.d. random variables $U_1, U_2, \ldots$ with $U_p \sim \chi^2_{u_p}$ such that

$$N\left(d_S(\theta,(Z_1,\ldots,Z_N)) - \frac{1}{2^q}\right) \xrightarrow{\mathcal{L}} \sum_{p=1}^{\infty} \frac{(q+1)!}{(q-1)!2^q}\frac{\lambda_p}{\tau_q}\left(U_p - u_p\right).$$

A simple possibility for estimating the quantiles is the generation of random numbers of the distribution. The quantiles given in Table 1 were calculated by computing 10 000 random numbers of the distribution (only the first 150 summands). The calculation of the quantiles was repeated 500 times. The means of these quantiles are given in the table. The 99.5% confidence band is ± 0.01 at most for each estimated quantile. The test statistic for multiple regression can be calculated similarly as for polynomial regression described in [20]. In particular it is shown how the test statistic can be calculated if the null hypothesis is a subspace of the parameter space or a polyhedron. But here the calculation of the simplicial depth of a given parameter is based on Lemma 1 in [17] by checking if $s_{n_1}(\theta)x_{n_1}$ is a linear combination of $s_{n_2}(\theta)x_{n_2}, \ldots, s_{n_{q+1}}(\theta)x_{n_{q+1}}$ with negative coefficients.

Table 1
Means of the simulated quantiles for multiple regression.

α-quantile (%)	$q = 2$	$q = 3$	$q = 4$
0.5	−2.607	−1.845	−1.222
1.0	−2.189	−1.566	−1.044
2.0	−1.771	−1.284	−0.863
2.5	−1.635	−1.192	−0.805
5.0	−1.216	−0.905	−0.619
10.0	−0.795	−0.612	−0.426
20.0	−0.368	−0.310	−0.224
30.0	−0.127	−0.126	−0.099
40.0	0.048	0.008	−0.006
50.0	0.183	0.116	0.072
60.0	0.293	0.209	0.140
70.0	0.388	0.292	0.203
80.0	0.473	0.373	0.265
90.0	0.554	0.456	0.331
95.0	0.600	0.504	0.373

4. Multiple regression with intercept

Proposition 1 showed that the asymptotic distribution of the simplicial depth does not depend on the unknown parameter if the distribution of the regressors does not depend on it, which is the case for usual regression models. But in general the asymptotic distribution depends on the underlying distribution of the explanatory variables. However, the next lemma shows for multiple regression with intercept that the asymptotic distribution does not depend on their location and scale.

Lemma 1. *Let $(Y_1, T_1, E_1), \ldots, (Y_N, T_N, E_N)$ be i.i.d random vectors such that there is a $\theta \in \mathbb{R}^q$ with*

$$Y_n = \theta_0 + \theta_1 T_{n,1} + \cdots + \theta_{q-1} T_{n,q-1} + E_n = x(T_n)^T \theta + E_n,$$

where $T_n = (T_{n,1}, \ldots, T_{n,q-1})$ and $X_n = x(T_n) = (1, T_{n,1}, \ldots, T_{n,q-1})^T$. Suppose that

- $P_\theta(Y_n - x(T_n)^T\theta > 0 | T_n) = \dfrac{1}{2}$
- $P_\theta(Y_n - x(T_n)^T\theta = 0 | T_n) = 0$
- $P_\theta(X_1, \ldots, X_q \text{ are linearly dependent}) = 0$
- $T_n = \mu + AV_n$ *for a* $\mu \in \mathbb{R}^{q-1}$, *an invertible matrix* $A \in \mathbb{R}^{q-1 \times q-1}$, *and a random vector* V_n.

Then the asymptotic distribution of the simplicial depth which is based on the dependent variable Y_n and the regressor X_n does not depend on μ and A.

The previous lemma shows that critical values for tests could be obtained by simulation if the general form of the distribution of the regressors is known. The exact form of the asymptotic distribution could be obtained for multivariate Cauchy distributed explanatory variables as follows:

We define two different statistical models with different simplicial depths. We want to calculate the asymptotic distribution of the simplicial depth d_S for a statistical model $(\mathcal{Z}^N, \mathcal{A}, \mathcal{P})$ with $\mathcal{P} = \{\otimes_{n=1}^N P_\theta : \theta \in \Theta\}$. We consider another statistical model $(\tilde{\mathcal{Z}}^N, \tilde{\mathcal{A}}, \tilde{\mathcal{P}})$ with $\tilde{\mathcal{P}} = \{\otimes_{n=1}^N \tilde{P}_\theta : \theta \in \Theta\}$ and for this model, we define a simplicial depth $\tilde{d}_S$. We show that the distribution of the simplicial depth d_S in the first model is equal to the distribution of the simplicial depth $\tilde{d}_S$ in the second model, i.e. $(\otimes_{n=1}^N P_\theta)^{d_S(\theta,\cdot)} = (\otimes_{n=1}^N \tilde{P}_\theta)^{\tilde{d}_S(\theta,\cdot)}$. To prove the next theorem, we introduce additional standard normal distributed random variables S_n so that the simplicial depth in the second model, which is a model for multiple regression through the origin, based on the dependent variable $S_n Y_n$ and the regressor $S_n x(T_n)$.

Theorem 2. *Let $(Y_1, T_1, E_1), \ldots, (Y_N, T_N, E_N)$ be i.i.d continuous distributed random vectors such that there is a $\theta \in \mathbb{R}^q$ with*

$$Y_n = \theta_0 + \theta_1 T_{n,1} + \cdots + \theta_{q-1} T_{n,q-1} + E_n = x(T_n)^T \theta + E_n,$$

where $T_n = (T_{n,1}, \ldots, T_{n,q-1})$ and $X_n = x(T_n) = (1, T_{n,1}, \ldots, T_{n,q-1})^T$. Suppose that

- $P_\theta(Y_n - x(T_n)^T\theta > 0 | T_n) = \dfrac{1}{2}$
- $P_\theta(Y_n - x(T_n)^T\theta = 0 | T_n) = 0$
- $f^{T_n}(t) = \dfrac{\Gamma(\frac{q}{2})}{\sqrt{\pi^q |\Sigma|}} \dfrac{1}{(1 + (t-\mu)^T \Sigma^{-1} (t-\mu))^{\frac{q}{2}}}.$

That is, T_n has a multivariate Cauchy Distribution. Let $Z_n = (Y_n, T_n)$. Then the asymptotic distribution of the simplicial depth which is based on the dependent variable Y_n and the regressor X_n is equal to the distribution given in Theorem 1.

5. Simulation studies

The assumption of Cauchy distributed regressors for the test for multiple regression with intercept may be only a technical requirement resulting from the proofs. In a simulation study we checked how the distribution of the test statistic depends on the distribution of the explanatory variables.

In the model

$$Y_n = \theta_0 + \theta_1 T_{n,1} + \theta_2 T_{n,2} + E_n, \tag{6}$$

we simulated the test statistic under the null hypothesis $H_0 : \theta = 0$. Because of Lemma 1 it suffices to use standardized distributions for the explanatory variables, at least for large sample sizes. Moreover, since the simplicial depth depends only on the signs of the residuals, the choice of the continuous and centered distribution of E_n does not affect the simulation results.

The observations were simulated under the null hypothesis with $E_n \sim \mathcal{N}(0, 1)$. We compared $T_n \sim \text{Cauchy}_2\left(\binom{0}{0}, I\right)$ with $T_n \sim \mathcal{N}_2\left(\binom{0}{0}, I\right)$, $f_{T_n} = \frac{2}{3} f_{\mathcal{N}_2\left(\binom{0}{0}, I\right)} + \frac{1}{3} f_{\mathcal{N}_2\left(\binom{5}{5}, 0.2I\right)}$, and $f_{T_n} = \frac{1}{4} f_{\mathcal{N}_2\left(\binom{0}{0}, 0.2I\right)} + \frac{1}{4} f_{\mathcal{N}_2\left(\binom{10}{0}, 0.2I\right)} + \frac{1}{4} f_{\mathcal{N}_2\left(\binom{0}{10}, 0.2I\right)} + \frac{1}{4} f_{\mathcal{N}_2\left(\binom{10}{10}, 0.2I\right)}$ respectively, where I is the identity matrix.

Realizations of the test statistic were obtained from 10 000 replications with sample size 50. The Kolmogorov–Smirnov tests could not reject the null hypotheses that the simulated test statistics have the same distribution as the test statistics simulated with $T_n \sim \text{Cauchy}_2\left(\binom{0}{0}, I\right)$ (P-values > 0.1). Hence it seems that the distribution of the explanatory variables has no influence on the asymptotic distribution of the test statistics although we have no proof for this. Note that the distribution of the test statistics in the finite case is very close to a continuous distribution.

In another simulation study, we tested the hypothesis $H_0 : \theta = 0$ at the level $\alpha = 0.05$ to compare the power of the simplicial depth test with the regression depth test according to Van Aelst et al. [27] and the F-test. We used the model for multiple regression with two explanatory variables given by Eq. (6), where $\theta = (0, 0, \theta_2)^T$ was the underlying parameter of the alternative.

In the first setting, data were simulated using Cauchy distributed explanatory variables $T_n \sim \text{Cauchy}_2\left(\binom{0}{0}, I\right)$ and Cauchy distributed errors $E_n \sim \text{Cauchy}(0, 1)$. Cauchy distributed explanatory variables simulate outliers in the explanatory variables and in particular leverage points. We chose Cauchy distributed errors in order to simulate outliers in the dependent variable as well. The critical value for the regression depth test was estimated only once according to the underlying distribution of the regressors and not separately for each realization of the regressors as proposed by Van Aelst et al. [27]. The power curves are obtained from 5000 repetitions. The results are shown in Fig. 1 for a sample size of 50. The F-test nearly keeps the level but has poor power for such observations. The probability to reject under the null hypothesis was slightly below 0.05 for the regression depth test because we did not randomize. The power of the simplicial depth test was slightly better than the power of the regression depth test.

In the second setting, data were simulated using normal distributed explanatory variables $T_n \sim \mathcal{N}_2\left(\binom{0}{0}, I\right)$ and normal distributed errors $E_n \sim \mathcal{N}(0, 1)$. The results are shown in Fig. 2. As expected, the F-test performed better than the others. Again, the power of the simplicial depth test was slightly better than the power of the regression depth test.

Moreover, the simulations show that the simplicial depth test works well for rather small sample sizes although it is an asymptotic test. Similar results were obtained in a power comparison for simple linear regression ($q = 2$) in [17].

6. Application: Test for multiple regression through the origin

The North American Sunfish "pumpkinseed" (*Lepomis gibbosus*) was introduced to European waters about 100 years ago. Near Brighton, 162 specimens were collected in 2003 from the Tanyards fisheries pond. Nineteen landmarks (see Fig. 3) were identified for each fish. The data is available by request from the authors.

In this section, we want to find out relationships between some of the landmarks. We restrict ourselves on those relationships, that can be tested within the model for multiple regression through the origin (for other ones, see e.g. [26] or [20]). We rotate, rescale and translate the fishes (the landmarks), such that landmark 10 (anterior tip of the upper jaw) is equal to $(-\frac{1}{2}, 0)^T$ and landmark 11 (caudal fin base) is equal to $(\frac{1}{2}, 0)^T$.

Let $\lambda_n^p = (\lambda_{n,1}^p, \lambda_{n,2}^p)^T \in \mathbb{R}^2$ be landmark number p of the nth transformed fish. Fig. 3 shows, that the horizontal position of the anterior edge of the dorsal fin base $\lambda_{n,1}^{19}$ is nearly equal to the horizontal position of the anterior edge of the pelvic fin base $\lambda_{n,1}^1$. We call

$$y_n = \lambda_{n,1}^{19} - \lambda_{n,1}^1$$

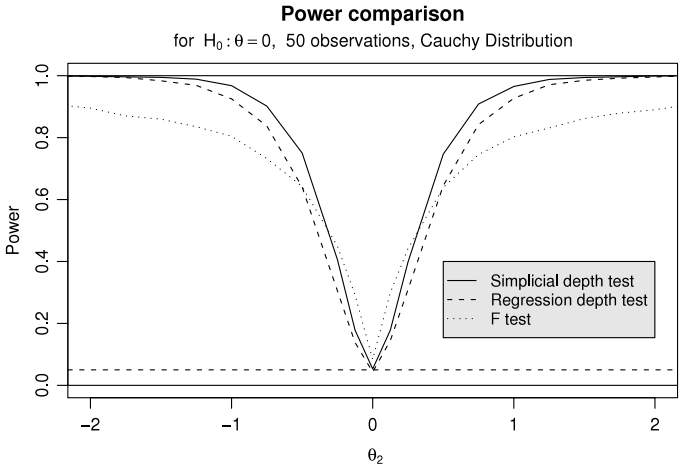

Fig. 1. Power comparison: Cauchy distribution.

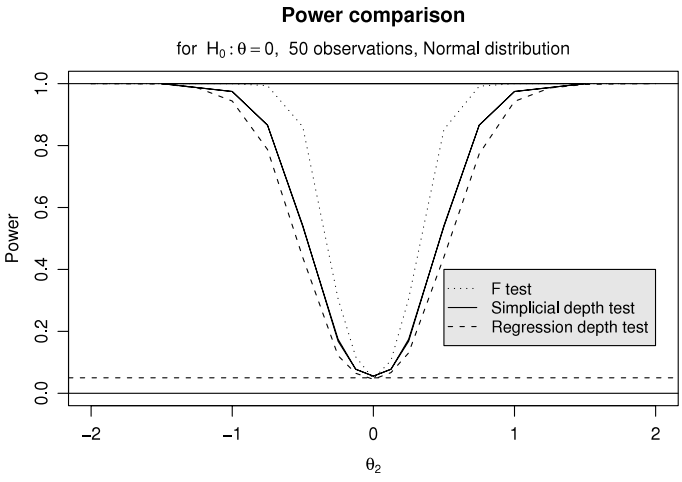

Fig. 2. Power comparison: Normal distribution.

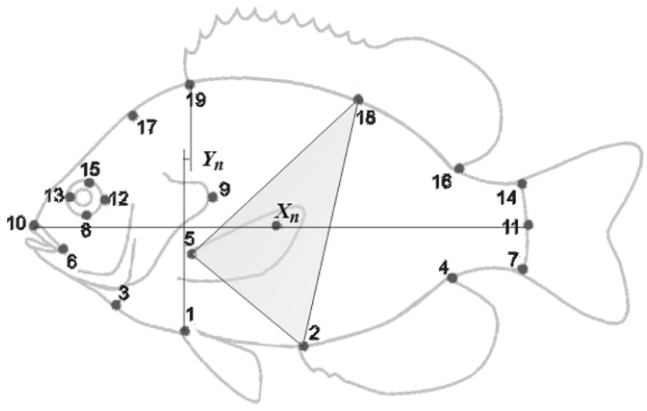

Fig. 3. Landmarks.

the fin base difference in this paper. The sign test for testing that the fin base difference is centered provides the very high p-value 0.937. We defined the center of the fish as a centered convex combination of 3 landmarks by the formula

$$x_n = 0.34\lambda_n^{18} + 0.22\lambda_n^{2} + 0.44\lambda_n^{5}.$$

Since the fin base difference and the center of the fish could both be influenced by the form of the vertebral column, there could be a dependency between Y_n and X_n. We test within the model for multiple regression through the origin ($q = 2$),

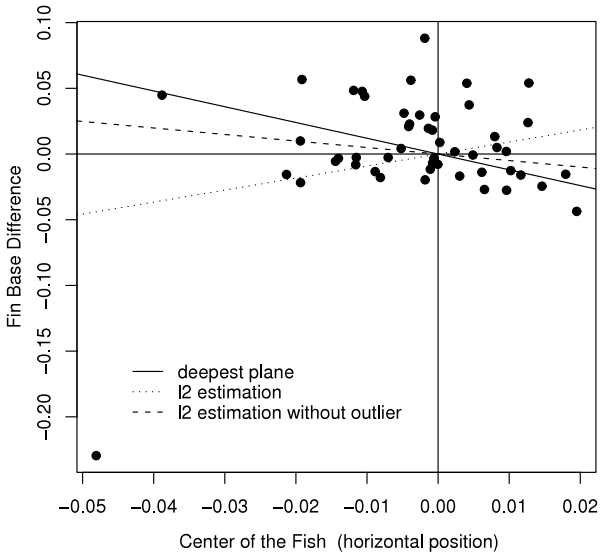

Fig. 4. A deepest plane with $\theta_2 = 0$ and least squares fits at the $x_{n,1}$-axis.

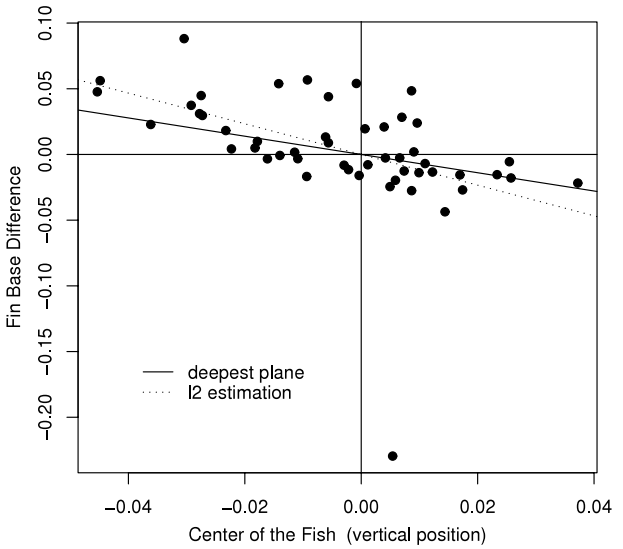

Fig. 5. A deepest plane with $\theta_1 = 0$ and the least square fit at the $x_{n,2}$-axis.

how Y_n depends on the center $X_n = (X_{n,1}, X_{n,2})^T$ of the fish. Therefore we choose a random sample that consists on 50 fishes. The original data are discrete, due to rounding errors. To make them continuous, we add a uniformly distributed random number in $[-0.005, 0.005)$ to each landmark so that we would obtain the original data by rounding.

The parameter with maximum simplicial depth is $\widehat{\theta}_D := (-0.541, -0.866)^T$ and the least squares fit is $\widehat{\theta}_{l_2} = (0.915, -1.168)^T$. At first we test the hypothesis, that $X_{n,1}$ has no influence on Y_n, that is, $H_0 : \theta_1 = 0$. The test statistic depends on the depth of the deepest plane with $\theta_1 = 0$, given by the parameter $(0, -0.695)^T$ (see Fig. 5). The value of the test statistic is 0.122, which is more than the 40% quantile of the asymptotic distribution and thus, we have no rejection (see Table 1). Hence, we may assume that Y_n does not depend on the horizontal position of the center. Contrary to this result, the classical F-test rejects this hypothesis with respect to a significance level 5% (p-value = 0.028). This is due to the outlier in the left lower corner of Fig. 4. The outlier strongly influences the first component of the least squares fit $\widehat{\theta}_{l_2}$, whose first component is positive (see the dashed line in Fig. 4).

Without the outlier, the least squares fit is $\tilde{\theta}_{l_2} := (-0.496, -0.963)^T$ so that its first component is negative. Then the classical F-test would not reject the null hypothesis with respect to a significance level 5%. Note that the least squares fit for the data without the outlier is close to the parameter $\widehat{\theta}_D$ with maximum simplicial depth.

On the other hand, the hypothesis that $X_{n,2}$ has no influence on Y_n, that is, $H_0 : \theta_2 = 0$ has to be rejected with respect to a significance level 2%, since the test statistic -2.184 is near the 1% quantile of the asymptotic distribution. In particular,

the deepest plane with $\theta_2 = 0$ given by the parameter $(-1.2, 0)^T$ does not give a good description of the data (see Fig. 4). The classical F-test also rejects the null hypothesis and provides a p-value of 0.0001. Indeed, the least square fit is strongly decreasing at the $x_{n,2}$-axis (see the dashed line in Fig. 5).

We conclude that Y_n depends on $X_{n,2}$, but not on $X_{n,1}$. As shown in Fig. 5, the fin base difference becomes smaller if the center of the fish is shifted upwards. Roughly speaking, λ_n^{19} shifts to the left and/or λ_n^1 shifts to the right, if the center is shifted upwards. This is possibly due to a curved vertebral column. If this interpretation is correct, then one could take into consideration a nonlinear transformation of the landmarks before further investigations, such that the vertebral columns of the transformed fishes can be expected to be a straight line.

7. Proofs

See also [21] for details of the proofs.

Proof of Proposition 1. Let $x_1, x_2 \in \mathbb{R}^q \setminus \{0\}$. For $j = 3, \ldots, q+1$ let $W_j := \frac{1}{\|AX_j\|} AX_j$, $W := W_3 \times \cdots \times W_{q+1}$, and $U := \frac{W}{\|W\|}$ and for $j = 1, 2$ let

$$K^+(x_j) := \{w \in \mathbb{R}^q : (Ax_j)^T w \geq 0\},$$
$$K^-(x_j) := \{w \in \mathbb{R}^q : (Ax_j)^T w \leq 0\}.$$

Then we have

$$\begin{aligned}
\mathcal{K}(x_1, x_2) + \frac{1}{2} &= P(x_1^T(X_3 \times \cdots \times X_{q+1})x_2^T(X_3 \times \cdots \times X_{q+1}) < 0) \\
&= P(\det(x_1, X_3, \ldots, X_{q+1}) \det(x_2, X_3, \ldots, X_{q+1}) < 0) \\
&= P(\det(A(x_1, X_3, \ldots, X_{q+1})) \det(A(x_2, X_3, \ldots, X_{q+1})) < 0) \\
&= P(\det(Ax_1, AX_3, \ldots, AX_{q+1}) \det(Ax_2, AX_3, \ldots, AX_{q+1}) < 0) \\
&= P(\det(Ax_1, W_3, \ldots, W_{q+1}) \det(Ax_2, W_3, \ldots, W_{q+1}) < 0) \\
&= P((Ax_1)^T U (Ax_2)^T U < 0) \\
&= P(U \in K^+(x_1) \cap K^-(x_2)) + P(U \in K^-(x_1) \cap K^+(x_2)) \\
&= P(U \in K^+(x_1) \cap K^-(x_2)) + P(-U \in K^+(x_1) \cap K^-(x_2)). \qquad (7)
\end{aligned}$$

We show that W is orthogonally invariant. Let $\Pi \in \mathbb{R}^{q \times q}$ be an orthogonal matrix and let $w_3, .., w_{q+1} \in \mathbb{R}^q$. For all $w \in \mathbb{R}^q$ we have

$$\begin{aligned}
(\Pi w_3 \times \cdots \times \Pi w_{q+1})^T \Pi w &= \det(\Pi w_3, \ldots, \Pi w_{q+1}, \Pi w) \\
&= \det(\Pi(w_3, \ldots, w_{q+1}, w)) \\
&= \det(\Pi) \det(w_3, \ldots, w_{q+1}, w) \\
&= \det(\Pi)(w_3 \times \cdots \times w_{q+1})^T w,
\end{aligned}$$

so that

$$(\Pi w_3 \times \cdots \times \Pi w_{q+1})^T \Pi = \det(\Pi)(w_3 \times \cdots \times w_{q+1})^T,$$

and

$$\Pi^T(\Pi w_3 \times \cdots \times \Pi w_{q+1}) = \det(\Pi)(w_3 \times \cdots \times w_{q+1}).$$

Since Π is an orthogonal matrix we have $\Pi^T = \Pi^{-1}$ and $\det(\Pi) \in \{-1, 1\}$. Moreover $W_3, \ldots, W_{q+1}$ are independent and orthogonally invariant. It follows for each event $B \subset \mathbb{R}^q$:

$$\begin{aligned}
P(W \in B) &= P(\Pi^T(W_3 \times \cdots \times W_{q+1}) \in \Pi^{-1}B) \\
&= P(\Pi^T(\Pi W_3 \times \cdots \times \Pi W_{q+1}) \in \Pi^{-1}B) \\
&= P(\det(\Pi)(W_3 \times \cdots \times W_{q+1}) \in \Pi^{-1}B) \\
&= P((\det(\Pi)W_3) \times \cdots \times W_{q+1} \in \Pi^{-1}B) \\
&= P(W \in \Pi^{-1}B) \\
&= P(\Pi W \in B),
\end{aligned}$$

so that W is orthogonally invariant. It follows with Devroye [22] that $U = \frac{W}{\|W\|}$ is uniformly distributed on the unity sphere, see also [23]. Because of $-U \sim U$, we obtain from Eq. (7) that

$$\mathcal{K}(x_1, x_2) = 2P(U \in K^+(x_1) \cap K^-(x_2)) - \frac{1}{2}.$$

The proportion of the unit sphere, that is contained in $K^+(x_1) \cap K^-(x_2)$ is equal to the angle between Ax_1 and Ax_2, divided by 2π.

Hence,

$$\begin{aligned}\mathcal{K}(x_1, x_2) &= 2\frac{\sphericalangle(Ax_1, Ax_2)}{2\pi} - \frac{1}{2} \\ &= \frac{1}{\pi}\arccos\left(\left\langle \frac{Ax_1}{\|Ax_1\|}, \frac{Ax_2}{\|Ax_2\|}\right\rangle\right) - \frac{1}{2}. \quad \square\end{aligned}$$

Proof of Proposition 2. Since the required Gegenbauer functions have different definitions for $q = 2$ and $q \geq 3$, both cases have to be handled separately. At first, we investigate the case $q \geq 3$.

For brevity let us write $\lambda := \frac{n}{2}$. For all $s, t \in S$ we have

$$\begin{aligned}K(s, t) &= \frac{1}{\pi}\arccos(\langle s, t\rangle) - \frac{1}{2} \\ &= \frac{1}{\pi}\arccos(\cos(\sphericalangle(s, t))) - \frac{1}{2} \\ &= k(\cos(\sphericalangle(s, t))),\end{aligned}$$

where $k(\sigma) := \frac{1}{\pi}\arccos(\sigma) - \frac{1}{2} \in \mathbb{L}^2[-1, 1]$.

Since the kernel function only depends on $\cos(\sphericalangle(s, t))$, it follows by Fenyö and Stolle [19, p. 273], that $\{S^{(n)}_{(p,l)}\}$ is the complete system of eigenfunctions of T_K with eigenvalues

$$\begin{aligned}\lambda_p &= \frac{4\pi^{\frac{n}{2}+1}}{(2p+n)\Gamma\left(\frac{n}{2}\right)} b_p c_p, \quad \text{for } p \in \mathbb{N}_0, \text{ where} \\ b_p &:= \frac{2^{n-1}p!\left(\frac{n}{2}+p\right)\Gamma\left(\frac{n}{2}\right)^2}{\pi\,\Gamma(n+p)}, \\ c_p &:= \int_{-1}^{1} k(\sigma) C_p^{\lambda}(\sigma)\left(1-\sigma^2\right)^{\frac{(n-1)}{2}} \mathrm{d}\sigma.\end{aligned}$$

We denote by C_p^{λ} the $(n+2)$-dimensional Gegenbauer function. Useful properties of this function are derived in [24].

Since $\lambda > 0$ we have

$$C_p^{\lambda}(x) = \frac{\prod_{j=0}^{p-1}(2\lambda + j)}{\prod_{j=0}^{p-1}(\lambda + \frac{1}{2} + j)} P_p^{\left(\lambda-\frac{1}{2}, \lambda-\frac{1}{2}\right)}(x),$$

where

$$P_p^{(\alpha,\beta)}(x) = \frac{1}{2^p}\sum_{k=0}^{p} \frac{\prod_{m=0}^{k-1}(p+\alpha-k+1+m) \cdot \prod_{m=0}^{p-k-1}(\beta+k+1+m)}{k!(p-k)!}(x-1)^{p-k}(x+1)^k$$

is a Jacobi polynomial. For instance, see [24, p. 161 and p. 178].

By the doubling formula of the Gamma function $\Gamma(2z) = \frac{2^{2z-1}}{\sqrt{\pi}}\Gamma(z)\Gamma(z+\frac{1}{2})$ we obtain:

$$\lambda_p = \pi^{\lambda} p \frac{\Gamma(\lambda)\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{p}{2}+\frac{1}{2}\right)}{\Gamma\left(\lambda+\frac{p}{2}\right)\Gamma\left(\lambda+\frac{p}{2}+\frac{1}{2}\right)} c_p.$$

Because of $C_0^{\lambda} \equiv 1$ and $\arcsin(-x) = -\arcsin(x)$ we obtain

$$\begin{aligned}c_0 &= \int_{-1}^{1} \frac{1}{\pi}\left(\arccos(x) - \frac{\pi}{2}\right)\left(1-x^2\right)^{\lambda-\frac{1}{2}} \mathrm{d}x \\ &= -\int_{-1}^{1} \frac{1}{\pi}\arcsin(x)\left(1-x^2\right)^{\lambda-\frac{1}{2}} \mathrm{d}x \\ &= -\int_{-1}^{0} \frac{1}{\pi}\arcsin(x)\left(1-x^2\right)^{\lambda-\frac{1}{2}} \mathrm{d}x - \int_{0}^{1} \frac{1}{\pi}\arcsin(x)\left(1-x^2\right)^{\lambda-\frac{1}{2}} \mathrm{d}x\end{aligned}$$

$$= -\int_0^1 \frac{1}{\pi}\arcsin(-x)\left(1-x^2\right)^{\lambda-\frac{1}{2}}\mathrm{d}x - \int_0^1 \frac{1}{\pi}\arcsin(x)\left(1-x^2\right)^{\lambda-\frac{1}{2}}\mathrm{d}x$$
$$= 0.$$

Hence, $\lambda_0 = 0$. The function

$$F(x) := -\frac{2\lambda}{p(p+2\lambda)}C_{p-1}^{\lambda+1}(x)\left(1-x^2\right)^{\frac{1}{2}+\lambda}$$

has the derivative (see [25], p. 220):

$$F'(x) = C_p^{\lambda}(x)\left(1-x^2\right)^{\lambda-\frac{1}{2}}.$$

This is needed to simplify c_p for $p > 0$.

Let $p > 0$. Since $\arccos'(x) = -\left(1-x^2\right)^{-\frac{1}{2}}$ we obtain by integration by parts and [24, p. 179]:

$$\begin{aligned}
c_p &= \int_{-1}^1 \left(\frac{1}{\pi}\arccos(x) - \frac{1}{2}\right) C_p^{\lambda}(x)\left(1-x^2\right)^{\lambda-\frac{1}{2}}\mathrm{d}x \\
&= \frac{1}{\pi}\int_{-1}^1 \arccos(x)C_p^{\lambda}(x)\left(1-x^2\right)^{\lambda-\frac{1}{2}}\mathrm{d}x - \frac{1}{2}\int_{-1}^1 C_p^{\lambda}(x)C_0^{\lambda}(x)\left(1-x^2\right)^{\lambda-\frac{1}{2}}\mathrm{d}x \\
&\overset{\text{Tricomi}}{=} \frac{1}{\pi}\int_{-1}^1 \arccos(x)C_p^{\lambda}(x)\left(1-x^2\right)^{\lambda-\frac{1}{2}}\mathrm{d}x - 0 \\
&= \frac{1}{\pi}\left(\left[F(x)\arccos(x)\right]_{-1}^1 - \int_{-1}^1 \arccos'(x)F(x)\mathrm{d}x\right) \\
&= \frac{1}{\pi}\left(0 - 0 + \int_{-1}^1 \left(1-x^2\right)^{-\frac{1}{2}}F(x)\mathrm{d}x\right) \\
&= -\frac{1}{\pi}\int_{-1}^1 \left(1-x^2\right)^{-\frac{1}{2}}\frac{2\lambda}{p(p+2\lambda)}C_{p-1}^{\lambda+1}(x)\left(1-x^2\right)^{\frac{1}{2}+\lambda}\mathrm{d}x \\
&= -\frac{2\lambda}{\pi p(p+2\lambda)}\int_{-1}^1 C_{p-1}^{\lambda+1}(x)\left(1-x^2\right)^{\lambda}\mathrm{d}x.
\end{aligned}$$

The calculation of this integral is somewhat tedious, so we give only the result:

$$\int_{-1}^1 C_{p-1}^{\lambda+1}(x)\left(1-x^2\right)^{\lambda}\mathrm{d}x = \frac{\Gamma\left(\frac{p}{2}+\lambda+\frac{1}{2}\right)}{\Gamma\left(\frac{p}{2}+\lambda+1\right)}\frac{\Gamma\left(\frac{p}{2}\right)}{\Gamma\left(\frac{p}{2}+\frac{1}{2}\right)}\sin\left(\frac{p}{2}\pi\right)^2.$$

Another (rather ugly) expression for this integral can easily be obtained by the explicit representation of $C_{p-1}^{\lambda+1}$. Note, that $\lambda + 1 = \frac{q}{2}$. Putting together all steps, we obtain:

$$\begin{aligned}
\lambda_p &= \pi^{\lambda}p\frac{\Gamma(\lambda)\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{p}{2}+\frac{1}{2}\right)}{\Gamma\left(\lambda+\frac{p}{2}\right)\Gamma\left(\lambda+\frac{p}{2}+\frac{1}{2}\right)}c_p \\
&= -\pi^{\lambda}p\frac{\Gamma(\lambda)\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{p}{2}+\frac{1}{2}\right)}{\Gamma\left(\lambda+\frac{p}{2}\right)\Gamma\left(\lambda+\frac{p}{2}+\frac{1}{2}\right)}\frac{\lambda}{\pi p\left(\lambda+\frac{p}{2}\right)}\int_{-1}^1 C_{p-1}^{\lambda+1}(x)\left(1-x^2\right)^{\lambda}\mathrm{d}x \\
&= -\pi^{\lambda-1}\frac{\Gamma(\lambda+1)\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{p}{2}+\frac{1}{2}\right)}{\Gamma\left(\lambda+\frac{p}{2}+1\right)\Gamma\left(\lambda+\frac{p}{2}+\frac{1}{2}\right)}\frac{\Gamma\left(\frac{p}{2}+\lambda+\frac{1}{2}\right)}{\Gamma\left(\frac{p}{2}+\lambda+1\right)}\frac{\Gamma\left(\frac{p}{2}\right)}{\Gamma\left(\frac{p}{2}+\frac{1}{2}\right)}\sin\left(\frac{p}{2}\pi\right)^2 \\
&= \frac{-\pi^{\lambda+1}}{\Gamma(\lambda+1)}\frac{\Gamma(\lambda+1)^2\Gamma\left(\frac{p}{2}\right)^2}{\Gamma\left(\lambda+\frac{p}{2}+1\right)^2}\frac{\sin\left(\frac{p}{2}\pi\right)^2}{\pi^2} \\
&= -\frac{1}{2}2\frac{\pi^{\frac{q}{2}}}{\Gamma\left(\frac{q}{2}\right)}\frac{\Gamma\left(\frac{q}{2}\right)^2\Gamma\left(\frac{p}{2}\right)^2}{\Gamma\left(\frac{q}{2}+\frac{p}{2}\right)^2}\frac{\sin\left(\frac{p}{2}\pi\right)^2}{\pi^2} \\
&= -\frac{1}{2}\tau_q\left(\frac{\Gamma\left(\frac{q}{2}\right)\Gamma\left(\frac{p}{2}\right)}{\Gamma\left(\frac{q}{2}+\frac{p}{2}\right)}\frac{\sin\left(\frac{p}{2}\pi\right)}{\pi}\right)^2.
\end{aligned}$$

Now let $q = 2$. The eigenvalues for $q = 2$ can be obtained by calculating the formula

$$\lambda_p = \int_0^{2\pi} \left(\frac{1}{\pi} \arccos(\cos(\sigma)) - \frac{1}{2} \right) \cos(p\sigma) \mathrm{d}\sigma,$$

given in [19]. It is not difficult to show, that $\lambda_0 = 0$ and for $p \in \mathbb{N}$ we have

$$\begin{aligned}
\lambda_p &= \int_0^{\pi} \left(\frac{\sigma}{\pi} - \frac{1}{2} \right) \cos(p\sigma) \mathrm{d}\sigma + \int_{\pi}^{2\pi} \left(\frac{2\pi - \sigma}{\pi} - \frac{1}{2} \right) \cos(p\sigma) \mathrm{d}\sigma \\
&= -\frac{1}{2} 2\pi \frac{2(1 - \cos(p\pi))}{p^2 \pi^2} \\
&= -\frac{1}{2} 2\pi \left(\frac{2}{p} \frac{\sin(\frac{p}{2}\pi)}{\pi} \right)^2.
\end{aligned}$$

In order to validate the last equation, note that $\sin(\frac{p}{2}\pi)^2$ is just an indicator function. Hence, the proposition holds also for $q = 2$. □

Proof of Lemma 1. With $Z_n := (Y_n, T_n)$ we have $v(Z_n) = x(T_n)$ and $h(Z_n) = Y_n$. Since the random variables $Z_1, \ldots, Z_N$ satisfy assumption (4), the asymptotic distribution of the simplicial depth depends only on the kernel $\mathcal{K}$ given in Eq. (5). Let $\tilde{W} := x(V_3) \times \cdots \times x(V_{q+1})$ and let

$$\tilde{\mathcal{K}}(x(v_1), x(v_2)) := P_\theta(x(v_1)^T \tilde{W} x(v_2)^T \tilde{W} < 0) - \frac{1}{2}, \quad \text{for } v_1, v_2 \in \mathbb{R}^{q-1}.$$

The matrix $B := \begin{pmatrix} 1 & 0 \\ \mu & A \end{pmatrix}$ is invertible and we have $x(T_n) = Bx(V_n)$. Thus,

$$\begin{aligned}
\mathcal{K}(x(t_1), x(t_2)) + \frac{1}{2} &= P(x(t_1)^T (X_3 \times \cdots \times X_{q+1}) x(t_2)^T (X_3 \times \cdots \times X_{q+1}) < 0) \\
&= P(\det(x(t_1), X_3, \ldots, X_{q+1}) \det(x(t_2), X_3, \ldots, X_{q+1}) < 0) \\
&= P(\det(Bx(v_1), Bx(V_3), \ldots, Bx(V_{q+1})) \det(Bx(v_2), Bx(V_3), \ldots, Bx(V_{q+1})) < 0) \\
&= P(\det(B(x(v_1), x(V_3), \ldots, x(V_{q+1}))) \det(B(x(v_2), x(V_3), \ldots, x(V_{q+1}))) < 0) \\
&= P(\det(B)^2 \det(x(v_1), x(V_3), \ldots, x(V_{q+1})) \det(x(v_2), x(V_3), \ldots, x(V_{q+1})) < 0) \\
&= P(x(v_1)^T (x(V_3) \times \cdots \times x(V_{q+1})) x(v_2)^T (x(V_3) \times \cdots \times x(V_{q+1})) < 0) \\
&= \tilde{\mathcal{K}}(x(v_1), x(v_2)) + \frac{1}{2}.
\end{aligned}$$

The kernel $\tilde{\mathcal{K}}$ does not depend on A and μ and the eigenvalues $\lambda_1, \lambda_2, \ldots \in \mathbb{R}$ from the spectral decomposition of $\tilde{\mathcal{K}}$ in $\mathbb{L}_2\left(P^{x(V_1)} \otimes P^{x(V_1)}\right)$ are identical to those of $\mathcal{K}$ in $\mathbb{L}_2\left(P^{X_1} \otimes P^{X_1}\right)$. Thus, the asymptotic distribution of the simplicial depth also does not depend on A and μ. □

Proof of Theorem 2. Because of Lemma 1 it suffices to prove the theorem for $\mu = 0$.

We compare the simplicial depth in the statistical model for $Z_1, \ldots, Z_N$ with a simplicial depth for i.i.d. random variables $\tilde{Z}_1, \ldots, \tilde{Z}_N$, where $\tilde{Z}_n$ is obtained from Z_n by appending an independent standard normal distributed random variable S_n. That is, $\tilde{Z}_n = (Z_n, S_n)$ and $\tilde{P}_\theta^{\tilde{Z}_n} := P_\theta^{Z_n} \otimes P_{\mathcal{N}(0,1)}$ is the distribution of $\tilde{Z}_n$. Take $\tilde{f}_\theta$ to be a density of $\tilde{P}_\theta^{\tilde{Z}_n}$.

Simplicial depth $\tilde{d}_S$ and tangent depth $\tilde{d}_T$ of θ with respect to the observations $\tilde{z}_n = (y_n, t_n, s_n)$ are based on the dependent variable $\tilde{h}(\tilde{z}_n) = s_n y_n$ and the regressor $\tilde{v}(\tilde{z}_n) = s_n x(t_n)$. Note, that the sign of the residual of observation $\tilde{z}_n = (z_n, s_n)$ is given by

$$\begin{aligned}
\tilde{\mathrm{sig}}_\theta(\tilde{z}_n) &= \mathrm{sign}\,(s_n y_n - s_n x(t_n)^T \theta) \\
&= \mathrm{sign}\,(s_n) \mathrm{sig}_\theta(z_n).
\end{aligned}$$

Since

$$\begin{aligned}
\tilde{d}_T(\theta, \tilde{z}) &= \min_{u \neq 0} \#\{\mathrm{sign}(s_n) \mathrm{sig}_\theta(z_n) s_n u^T x(t_n) > 0\} \\
&= \min_{u \neq 0} \#\{\mathrm{sig}_\theta(z_n) u^T x(t_n) > 0\} \\
&= d_T(\theta, z),
\end{aligned}$$

tangent depths are equal in both models for $s_1, \ldots, s_N \neq 0$. This holds also for the harmonized depths and thus, also the simplicial depths coincide, that is, for all $\theta \in \Theta$ and all $\tilde{z}_n = (z_n, s_n) \in \mathcal{Z} \times \mathbb{R}$ with $s_n \neq 0$ for $n = 1, \ldots, N$, we have

$$d_S(\theta, z) = \tilde{d}_S(\theta, \tilde{z}).$$

Thus,

$$\begin{aligned}\otimes_{n=1}^{N}\tilde{P}_{\theta}^{\tilde{Z}_n}(\{\tilde{z}:\tilde{d}_S(\theta,\tilde{z})<\lambda\}) &= (\otimes_{n=1}^{N}P_{\theta}^{Z_n})\otimes(\otimes_{n=1}^{N}P_{\mathcal{N}(0,1)})(\{z:d_S(\theta,z)<\lambda\}\times\mathbb{R}^N)\\ &= \otimes_{n=1}^{N}P_{\theta}^{Z_n}(\{z:d_S(\theta,z)<\lambda\})\end{aligned}$$

for all $\lambda>0$, so that also the distributions of the simplicial depths are equal in both models.

It remains to show that $\tilde{Z}_1,\ldots,\tilde{Z}_N$ satisfy the assumptions of Theorem 1. Since the random variables are continuously distributed, conditional densities can be used to check that $\tilde{\text{sig}}_{\theta}(\tilde{Z}_n)$ is positive (negative) with probability $\frac{1}{2}$, given $\tilde{v}(\tilde{Z}_n)=S_n x(T_n)$.

The main part is to show that $K(\tilde{Z}_1):=\frac{1}{\|A\tilde{v}(\tilde{Z}_1)\|}A\tilde{v}(\tilde{Z}_1)$ with $A=\begin{pmatrix}1 & 0\\ 0 & \Sigma^{-\frac{1}{2}}\end{pmatrix}$ is uniformly distributed on the unit sphere S. The random variable $U(y_1,t_1,s_1):=\Sigma^{-\frac{1}{2}}t_1$ is multivariate Cauchy distributed with density

$$\tilde{f}_{\theta}^{U}(u)=\frac{\Gamma(\frac{q}{2})}{\sqrt{\pi^q}}\frac{1}{(1+u^Tu)^{\frac{q}{2}}}$$

and for $\tilde{Z}_1=(Y_1,T_1,S_1)$ we can write

$$K(\tilde{Z}_1)=\text{sign}(S_1)\frac{1}{\sqrt{1+U(\tilde{Z}_1)^TU(\tilde{Z}_1)}}\begin{pmatrix}1\\ U(\tilde{Z}_1)\end{pmatrix}.$$

It suffices to show that

$$\frac{\mu(V)}{\mu(S)}=\int_V 1\mathrm{d}(\tilde{P}_{\theta}^{\tilde{Z}_1})^K$$

for each event $V\subset S\cap\mathbb{R}_{>0}\times\mathbb{R}^{q-1}$ and each event $V\subset S\cap\mathbb{R}_{<0}\times\mathbb{R}^{q-1}$ which is open in S, where μ is the uniform measure on S with $\mu(S)=\tau_q$.

Consider the case $V\subset S\cap\mathbb{R}_{>0}\times\mathbb{R}^{q-1}$. Letting

$$U(V):=\left\{u\in\mathbb{R}^{q-1}:\frac{1}{\sqrt{1+u^Tu}}(1,u_1,\ldots,u_{q-1})^T\in V\right\},$$

the function

$$\psi:U(V)\to V,\quad \psi(u):=\frac{1}{\sqrt{1+u^Tu}}((-1)^{i+1},u_1,\ldots,u_{q-1})^T$$

is a local parametrization of V. Hence,

$$\mu(V)=\int_{U(V)}\sqrt{g\psi(u)}\mathrm{d}\lambda^{q-1},$$

where the gram determinant $g\psi(u)$ is defined as

$$g\psi(u)=\det\begin{pmatrix}\sum_{j=1}^{q}\frac{\partial\psi_j}{\partial u_1}(u)\frac{\partial\psi_j}{\partial u_1}(u) & \ldots & \sum_{j=1}^{q}\frac{\partial\psi_j}{\partial u_1}(u)\frac{\partial\psi_j}{\partial u_{q-1}}(u)\\ \vdots & & \vdots\\ \sum_{j=1}^{q}\frac{\partial\psi_j}{\partial u_{q-1}}(u)\frac{\partial\psi_j}{\partial u_1}(u) & \ldots & \sum_{j=1}^{q}\frac{\partial\psi_j}{\partial u_{q-1}}(u)\frac{\partial\psi_j}{\partial u_{q-1}}(u)\end{pmatrix}.$$

It is tedious to check that

$$g\psi(u)=\frac{1}{(1+u^Tu)^{2(q-1)}}\det((1+u^Tu)I-uu^T),$$

where $I=(e_1,\ldots,e_{q-1})$ is the identity matrix. With $a_{1,j}:=(1+u^Tu)e_j$, and $a_{2,j}:=-u_ju$ for $j=1,\ldots,q-1$ we have

$$\det((1+u^Tu)I-uu^T)=\det(a_{1,1}+a_{2,1},\ldots,a_{1,q-1}+a_{2,q-1}).$$

Since the determinant is linear in each column and since the determinant of a matrix is 0, if two columns are linearly dependent, we obtain

$$\det((1+u^Tu)I-uu^T)=\det((1+u^Tu)I)+\sum_{i=1}^{q-1}(1+u^Tu)^{q-2}(-u_i^2)=(1+u^Tu)^{q-2}.$$

It follows that $g\psi(u) = \frac{1}{(1+u^Tu)^q}$ and thus,

$$\mu(V) = \int_{U(V)} \sqrt{\frac{1}{(1+u^Tu)^q}} \mathrm{d}\lambda^{q-1} \quad \text{for } V \subset S \cap \mathbb{R}_{>0} \times \mathbb{R}^{q-1}. \tag{1}$$

Now let $V \subset S \cap \mathbb{R}_{>0} \times \mathbb{R}^{q-1}$ or $V \subset S \cap \mathbb{R}_{<0} \times \mathbb{R}^{q-1}$ be open in S. Let $i = 1$ or $i = 2$, such that $(-1)^iV \subset S \cap \mathbb{R}_{>0} \times \mathbb{R}^{q-1}$. For brevity we write $\tilde{P}_\theta := \tilde{P}_\theta^{\tilde{Z}_1}$.

With $\bar{T}(y_n, t_n, s_n) := t_n$ and $\bar{S}(y_n, t_n, s_n) := s_n$, we have

$$\begin{aligned}
\int_V 1 d\tilde{P}_\theta^K &= \tilde{P}_\theta(K \in V) \\
&= \tilde{P}_\theta(K \in V | \bar{S} > 0)\tilde{P}_\theta(\bar{S} > 0) + \tilde{P}_\theta(K \in V | \bar{S} < 0)\tilde{P}_\theta(\bar{S} < 0) \\
&= \tilde{P}_\theta\left(\frac{Ax(\bar{T})}{\|Ax(\bar{T})\|} \in V\right)\frac{1}{2} + \tilde{P}_\theta\left(-\frac{Ax(\bar{T})}{\|Ax(\bar{T})\|} \in V\right)\frac{1}{2} \\
&= \frac{1}{2}\tilde{P}_\theta\left(\frac{Ax(\bar{T})}{\|Ax(\bar{T})\|} \in (-1)^iV\right) \\
&= \frac{1}{2}\tilde{P}_\theta\left(\psi^{-1}\left(\frac{1}{\sqrt{1+U^TU}}\begin{pmatrix}1\\U\end{pmatrix}\right) \in \psi^{-1}((-1)^iV)\right) \\
&= \frac{1}{2}\tilde{P}_\theta(U \in \psi^{-1}((-1)^iV)) \\
&= \frac{1}{2}\int_{\psi^{-1}((-1)^iV)} \tilde{f}_\theta^U(u)\mathrm{d}\lambda^{q-1} \\
&= \frac{\Gamma(\frac{q}{2})}{2\sqrt{\pi^q}}\int_{\psi^{-1}((-1)^iV)} \frac{1}{(1+u^Tu)^{\frac{q}{2}}}\mathrm{d}\lambda^{q-1} \\
&\overset{(1)}{=} \frac{\mu((-1)^iV)}{\mu(S)} = \frac{\mu(V)}{\mu(S)}.
\end{aligned}$$

It follows that the assumptions of Theorem 1 hold, so that $\tilde{d}_S$ has the asymptotic distribution, mentioned there. Since the distributions of the simplicial depths are equal, it follows that also d_S has that asymptotic distribution. □

Acknowledgments

The shape analysis data are from a study funded by the Slovak Scientific Grant Agency, Project No. 1/9113/02. The data base for the example was created by Stanislav Katina. We thank him for permission to include it. We are also very grateful for the suggestions and hints of the referees which improve the presentation of the paper. The second author's research was supported by the SFB/TR TRR 30 Project D6.

References

[1] R.Y. Liu, On a notion of simplicial depth, Proc. Nat. Acad. Sci. USA 85 (1988) 1732–1734.
[2] R.Y. Liu, On a notion of data depth based on random simplices, Ann. Statist. 18 (1990) 405–414.
[3] J.W. Tukey, Mathematics and the picturing of data, in: Proc. International Congress of Mathematicians, Vancouver, 1974, vol. 2, 1975, pp. 523–531.
[4] K. Mosler, Multivariate Dispersion, Central Regions and Depth. The Lift Zonoid Approach, in: Lecture Notes in Statistics, vol. 165, Springer, New York, 2002.
[5] P.J. Rousseeuw, M. Hubert, Regression depth (with discussion), J. Amer. Statist. Assoc. 94 (1999) 388–433.
[6] Y. Zuo, R. Serfling, General notions of statistical depth function, Ann. Statist. 28 (2000) 461–482.
[7] Y. Zuo, R. Serfling, Structural properties and convergence results for contours of sample statistical depth functions, Ann. Statist. 28 (2000) 483–499.
[8] I. Mizera, On depth and deep points: A calculus, Ann. Statist. 30 (2002) 1681–1736.
[9] I. Mizera, Ch.H. Müller, Location-scale depth, J. Amer. Statist. Assoc. 99 (2004) 949–966. With discussion.
[10] Ch.H. Müller, Depth estimators and tests based on the likelihood principle with application to regression, J. Multivariate Anal. 95 (2005) 153–181.
[11] R.Y. Liu, Data Depth and Multivariate Rank Tests, in: Y. Dodge (Ed.), L_1-Statistical Analysis and Related Methods, North-Holland, Amsterdam, 1992, pp. 279–294.
[12] R.Y. Liu, K. Singh, A quality index based on data depth and multivariate rank tests, J. Amer. Statist. Assoc. 88 (1993) 252–260.
[13] Y. Zuo, X. He, On the limiting distributions of multivariate depth-based rank sum statistics and related tests, Ann. Statist. 34 (2006) 2879–2896.
[14] H. Oja, Descriptive statistics for multivariate distributions, Statist. Probab. Lett. 1 (1983) 327–332.
[15] H. Oja, Affine invariant multivariate sign and rank tests and corresponding estimates: A review, Scand. J. Statist. 26 (1999) 319–343.
[16] Z.-D. Bai, X. He, Asymptotic distributions of the maximal depth estimators for regression and multivariate location, Ann. Statist. 27 (1999) 1616–1637.
[17] R. Wellmann, P. Harmand, Ch.H. Müller, Distribution free tests for polynomial regression based on simplicial depth, J. Multivariate Anal. 100 (2009) 622–635.
[18] A.J. Lee, *U*-Statistics. Theory and Practice, Marcel Dekker, New York, 1990.
[19] S. Fenyö, H.W. Stolle, Theorie und Praxis der linearen Integralgleichungen, vol. 2, Birkhäuser Verlag, Basel, 1983.

[20] R. Wellmann, S. Katina, Ch.H. Müller, Calculation of simplicial depth estimators for polynomial regression with applications, Comput. Statist. Data Anal. 51 (2007) 5025–5040.
[21] R. Wellmann, On data depth with application to regression and tests, Ph.D. Thesis, University of Kassel, Germany, 2007.
[22] L. Devroye, Non-Uniform Random Variate Generation, Springer-Verlag, New York, 1986.
[23] C. Vignat, A. Plastino, The p-sphere and the geometric substratum of power-law probability distributions, Phys. Lett. A 343 (2005) 411–416.
[24] F.G. Tricomi, Vorlesungen über Orthogonalreihen, Springer, Berlin, 1955.
[25] W. Magnus, F. Oberhettinger, R.P. Soni, Formulas and Theorems for the Special Functions of Mathematical Physics, Springer, Berlin, 1966.
[26] J. Tomeček, V. Kováč, S. Katina, Ontogenetic variability in external morphology of native (Canadian) and nonnative (Slovak) populations of pumpkinseed (Lepomis gibbosus, Linnaeus 1758), J. Appl. Ichthyology 21 (2005) 335–344.
[27] S. Van Aelst, P.J. Rousseeuw, M. Hubert, A. Struyf, The deepest regression method, J. Multivariate Anal. 81 (2002) 138–166.
[28] H. Witting, U. Müller-Funk, Mathematische Statistik II, Teubner, Stuttgart, 1995.